AD 400 569

REPORT NO. 1

30 OCTOBER 1962

400 569

RESEARCH IN INFORMATION RETRIEVAL

First Quarterly Report

1 July 1962 - 30 September 1962

Contract No. DA 36-039-SC-90787
File No. 1160-PM-62-93-93(6509)

Technical Report P-AA-TR-(0010)

U. S. Army Electronics Research and Development Laboratory
Fort Monmouth, New Jersey

International Electric Corporation
Route 17 and Garden State Parkway, Paramus, New Jersey
A SUBSIDIARY OF INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION

Report No. 1 30 October 1962

RESEARCH IN INFORMATION RETRIEVAL

First Quarterly Report
1 July 1962 - 30 September 1962

An investigation of the techniques and concepts of information retrieval

Contract No. DA 36-039-SC-90787
File No. 1160-PM-62-93-93(6509)

Signal Corps Technical Requirement
SCL-4218 12 January 1960

Technical Report P-AA-TR-(0010)

Jacques Harlow, Principal Investigator

prepared by
Alfred Trachtenberg
Quentin A. Darmstadt
George Greenberg

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. PURPOSE

1.1 SCOPE

This report discusses the work performed for the U. S. Army Signal Research and Development Laboratory under Contract No. DA 36-039-SC-90787 during the period from 1 July 1962 to 30 September 1962.

1.2 OBJECTIVES

The objective of this project is to investigate the techniques and concepts of information retrieval and to formulate and develop a general theory of information retrieval. The formalization of this theory is oriented to the automation of large-capacity information storage and retrieval systems. This theoretical framework will be the basis for designing a general purpose stored-program digital computer system to perform the storage and retrieval functions.

1.3 PROJECT TASKS

The initial phase of this period was spent in defining the frame of reference, including the limitations and constraints, for this research activity. During this phase activity was oriented to the possibility of defining an aspect of equipment design that could be fruitfully exploited. The difficulty encountered in this approach was the lack of a definitive theoretical concept to use as a foundation for design criteria.

The second phase was spent in evaluating existing storage and retrieval systems, primarily to discern the major functional characteristics of these systems. As a result, a limited number of basic characteristics, each with a small number of variations, were isolated. Subsequent activity has been

expended in analyzing the nature of these characteristics in terms of rudimentary information-store, interrogate-retrieve interrelationships.

In the third phase three tasks were defined, and activity was concentrated in these areas:

(a) Formulation of General Principles.
(b) Development of Information-Retrieval Model.
(c) Development of Functional Elements.

Activity will continue in these tasks, particularly as discussed in Section 4 of this report.

2. ABSTRACT

This report discusses research activity performed in the investigation of the techniques and concepts of information retrieval. The general problems of information storage and retrieval are reviewed to establish a framework for the development of general theoretical principles. A preliminary model is presented as a medium for analyzing the functional characteristics of the storage and retrieval problem. Specific aspects of the problem--descriptor systems, file structures, and search procedures--are examined; and several measures of relevance are evaluated.

3. PUBLICATIONS, REPORTS, AND CONFERENCES

3.1 TECHNICAL NOTES

The following internal technical memoranda were issued during this reporting period:

(a) IEC TECHNICAL NOTE, File No. P-AA-TN-(0033)-N, 16 July 1962; Recommendations for Research in Information Retrieval, Quentin A. Darmstadt and Alfred Trachtenberg.

(b) IEC TECHNICAL NOTE, File No. P-AA-TN-(0035)-N, 10 August 1962; Review of Present Day Information Retrieval Systems, Alfred Trachtenberg.

3.2 REPORTS

The following reports were issued during this reporting period:

(a) MONTHLY LETTER REPORT NO. 1, 1 July 1962 - 31 July 1962, File No. P-AA-TR-(0006), 3 August 1962; Research in Information Retrieval, Alfred Trachtenberg.

(b) MONTHLY LETTER REPORT NO. 2, 1 August 1962 - 31 August 1962, File No. P-AA-TR-(0009), 31 August 1962; Research in Information Retrieval, Alfred Trachtenberg.

3.3 CONFERENCES

The following conferences were held between IEC personnel and the Signal Corps:

(a) 5 July 1962--Meeting at IEC. Discussions of objectives and plans for the research activity were initiated. The formulation of a method of approach was requested for presentation at the next meeting.

(b) 17 July 1962--Meeting at IEC. The memorandum referenced in Paragraph 3.1(a) was used as the basis of discussions pertaining to the scope, development phases, alternative plans, and recommended direction for the project.

(c) 18 July 1962--Meeting at IEC. Informal discussion of Signal Corps objectives and goals for research activity.

(d) 9 August 1962--Meeting at Fort Monmouth, New Jersey. The memorandum referenced in Paragraph 3.1(b) was used as the basis of discussions pertaining to functional characteristics of information retrieval systems. No particular area of activity was selected for further study.

(e) 10 September 1962--Meeting at IEC. Several methods of relating descriptor systems in a generalized sense were discussed in relation to the requirements for a file structure. The analysis and development of a general theory was recommended as the objective of the project.

4. FACTUAL DATA

4.1 STATEMENT OF THE PROBLEM

4.1.1 Original Formulation - The technical requirement of the Signal Corps, as specified in SCL-4355, is for "...a research investigation of techniques and concepts necessary for the efficient mechanization of large-capacity information storage and retrieval systems." Among the future applied objectives suggested as guides for such research are "...problems of military significance; i.e., personnel files, intelligence data, etc."

4.1.2 Alternative Approaches - This statement of the problem leads to many alternative approaches that any specific research program may take. Some of the possibilities that arise, and that have been taken in the past, may be characterized as dichotomies:

(a) System oriented versus specific operation oriented.
(b) A real system (problem) versus a hypothetical system (problem).
(c) Hardware emphasis versus software emphasis.
(d) Reduction to canonic forms versus manipulation of canonic forms.

These dichotomies should not be construed as mutually exclusive alternatives from which one alternative must be chosen in each instance in order to define the research task. The following discussion explicates some implications of emphasizing certain approaches to the program and establishes the validity of de-emphasizing others.

4.1.2.1 System Oriented versus Specified Operation Oriented - The need for information retrieval arises whenever an individual has a question that he believes can be answered by referencing some pool of data;

for the present neither question nor answer is rigorously defined. In general, however, the concern of the user of an information retrieval system is not with any specific documentation processes but with obtaining the information required by his question.

In this sense most current information retrieval systems--except for those like Baseball (4) or ACSIMATIC (11)--are misnamed. They are only parts of a larger system containing many implicit operations performed by the user; and these operations are not even formally specified nor readily specifiable.

A system orientation to research on information retrieval would focus on the job of providing the answers to certain kinds of questions about certain kinds of data. Different job contexts (personnel selection, scientific research, or intelligence analysis) deal with different kinds of questions and different kinds of raw data organization. As a consequence, each job generally results in quite different operating systems if optimumly designed.

A specific operation orientation to research on information retrieval might justifiably ignore large aspects of a user's job problems and concentrate upon improving specific operations used in many kinds of information retrieval systems--descriptor assignment, index organization, or search procedure. Such research might deal with a spectrum of increasingly sophisticated approaches to specific information retrieval procedures. In the ideal case less sophisticated procedures might be special cases of more inclusive systematic or theoretical formulations.

There is an important asymmetry to be considered in selecting between these research orientations. The system orientation is directed to the optimum use of the state-of-the-art in doing a particular job. To the extent that state-of-the-art improvements are important in doing the system job, some of the research effort may also be directed to developing improved retrieval procedures. The procedure oriented approach is concerned with improvements in the state-of-the-art and need not concern itself with the specialized problems of any given information retrieval system. It may be tempting to select the system oriented strategy in the hope that unusual success may lead to state-of-the-art improvements; but even if no breakthroughs occur, at least a usable system will result.

4.1.2.2 Real versus Hypothetical Problem - The problems for research may be to design a system for a specific user possessing certain operational requirements or to develop a procedure for a specific existing information retrieval system. These alternatives are instances of the system oriented or specific procedure oriented approaches to a real problem, respectively. In contrast, work may proceed on the development of a hypothetical system or the refinement of a procedure without reference to a real system.

This dichotomy has been stated independently of the system versus procedure alternative. In practice, however, it is more prudent to adopt a procedure oriented research strategy in the absence of specific user requirements. If there are no user requirements, then, in order to maintain an artificial system orientation, energy must be diverted to the detailed specification of hypothetical system requirements that are

virtually certain never to coincide with any specific real job.

4.1.2.3 Hardware versus Software Emphasis - This distinction is generally familiar and requires no further definition. It is not independent of the preceding dichotomies. To the extent that research pertains to a real system, it is impossible to avoid detailed hardware considerations. To the extent that a more theoretical, procedure oriented study is being undertaken, hardware may become a secondary consideration for future development. However, procedure oriented research in regard to specific hardware may also be meaningful.

4.1.2.4 Reduction to Canonic Forms versus Manipulation of Canonic Forms - In any existing automated information retrieval system either data or question inputs (and, except for Baseball, both question and data inputs) must be highly restricted in canonic form or format. The selection of convenient canonic forms or formats for specific jobs requires creative system analysis and a system orientation. There are information retrieval system research and development programs such as the ACSIMATIC intelligence system or the Western Reserve Library system (9) whose major value (or shortcoming) is based upon the specification of a new information format for a specific job. Similarly there are procedure oriented studies focusing either upon the efficient manipulation of a specific canonic form--e.g., the multi-list processing techniques of Prywes, Gray, et al (2,3) for manipulating data in attribute-value form--or upon the automatic reduction of ordinary discourse to canonic form for automatic information retrieval--the only example of this case is Baseball.

4.1.3 Alternatives Selected - The original IEC position was relatively open with regard to these alternatives. It was assumed, however, that specific user requirements related to an eventual application of the present research might be available. Then a major aspect of a sophisticated automated system would involve the automated reduction of both questions and data to canonic form; this type of system would, therefore, require linguistic analysis. Both of these orientations have been de-emphasized in the discussions of project objectives. The only alternative among the dichotomies that has been clearly rejected, however, is work on a "real" system. While hardware considerations may thus remain secondary, at least during the early stages of the present program, it is desirable not to restrict the project orientation to any specific retrieval procedures at this time. Even on the question of reduction to canonical form, the only area that has been eliminated from consideration is extensive work on linguistic analysis rather than on more general problems such as methods of descriptor assignment.

4.1.4 Refined Statement of the Problem - The problem as presently conceived is to develop a general theory of information retrieval whose primary goal is its use as a system tool for the optimum design of specific information retrieval systems in the future. In terms of the dichotomies, the orientation is more to procedures rather than systems, to the hypothetical rather than the real, and to software rather than hardware. To the extent that language analysis is de-emphasized, the orientation is to the selection and manipulation of canonic forms rather than to the automatic conversion of ordinary discourse to canonic forms.

In no case, however, has an extreme pole of the dichotomies been selected. Thus, the orientation is clearly to a theory of systems that can be applied to the design of specific job oriented systems in their entirety rather than to a specific procedure(s) that may be valuable; to dealing with real contexts that may be of future interest to the Signal Corps, wherever possible, rather than necessarily limiting the study to abstract formalism; to the consideration of optimum hardware once software at the level of algorithm rather than machine code has been specified; and to the problem of conversion to canonic form when linguistic complexity is not the critical problem.

The following sections describe two aspects of the approach to this general problem. One is the formulation of a general model of the information retrieval process. The other is the selection of specific problems of procedure and technique; the only example thus far is the problem of relevance and its measurement. It is expected that the information retrieval model will provide a framework both for understanding the critical features of information retrieval systems of different levels of sophistication and for isolating critical areas of information retrieval procedures and techniques to focus upon for further developments.

4.2 SYSTEM MODELS

4.2.1 General - The name Information Retrieval System has been applied to a large number of systems of varying purpose and capability, from personnel file and literature search systems to systems that retrieve specific bits of information upon request. Outwardly, these systems seem to operate on different principles; but if a general information retrieval

theory is possible, it must be able to show the basic similarities of these systems. It is necessary, then, to examine the operation of each type of system in order to develop a model that would be valid for all systems.

Intuitively, the literature search problem with descriptor association is a form of content retrieval or an intermediate step toward it. At least theoretically, the process is continuous in the sense that complete content retrieval is a limiting case of document retrieval. In the following paragraphs this intuitive argument is developed more rigorously within the framework of a general retrieval model.

4.2.2 Formulation of General Retrieval Model - The basis of the classic literature search problem is: a collection of documents exists, and a researcher desires to select from this collection a document or documents that are pertinent to his interests. The usual approach to this problem has been an attempt to describe the stored documents by a small number of words or symbols and then to search through these words or symbols until some match has been obtained with a description of the area of interest. This process is illustrated in Figure 1.

Documents, I, are entered into the system and analyzed. On the basis of this analysis descriptors (i.e., terms identifying the nature of the document, i) are assigned, including a unique identifying number or address. This analysis and descriptor assignment has traditionally been performed by human beings, although methods for automatic descriptor assignment have been proposed (6,7).

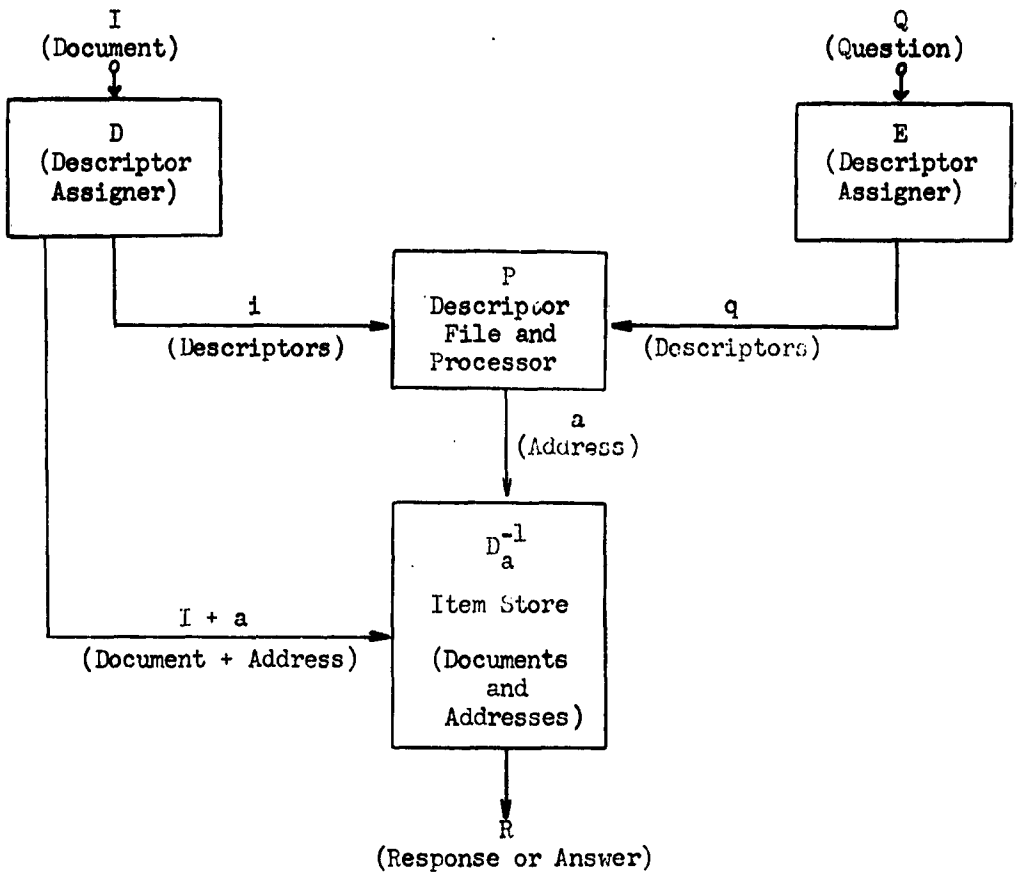


FIGURE 1. Model of Literature Search System

After the descriptors have been assigned, documents are placed in the item store in accordance with their assigned address. The complete set of descriptors is deposited in the descriptor file in accordance with the organization of the particular file. In the case of a library the descriptor file is the card catalogue, and the item store is simply the shelves on which the documents are stored.

A request for literature is translated into a set of descriptors

comparable to those that exist in the descriptor file. These descriptors are then matched against the descriptors in the descriptor file; when close enough matches have occurred, the addresses associated with these matches are noted and used to locate the desired items.

This process may be written in symbolic form in terms of the symbols in Figure 1 (or Figure 2):

(a) Documents are described:

$$i = D(I) \tag{4-1}$$

where i is a set of descriptors; I, a document; and D, a transformation algorithm.

(b) Questions are posed:

$$q = E(Q) \tag{4-2}$$

where q is a set of descriptors; Q, a question; and E, a transformation algorithm.

(c) Question and document descriptors are matched:

$$a = P(q,S_i) \tag{4-3}$$

where a is a set of unique addresses, which may be called an additional set of descriptors; S_i is the set of all descriptors in the descriptor file; and P is a transformation algorithm.

(d) The desired documents are located:

$$R = D_a^{-1}(a) \tag{4-4}$$

where D_a^{-1} is the inverse of the address assigning transformation algorithm. This function might also be written more generally as $R = D^{-1}(a)$; for D_a is part of D (see Equation 4-5).

An inherent difficulty of existing literature search systems is that the response may include superfluous information. Now, let content

retrieval be defined as the process for obtaining specific information, accompanied by little or no superfluous information, in response to a query. If Figure 1 is then restructured as shown in Figure 2, it is clear that this primitive model is valid for literature search and such lesser

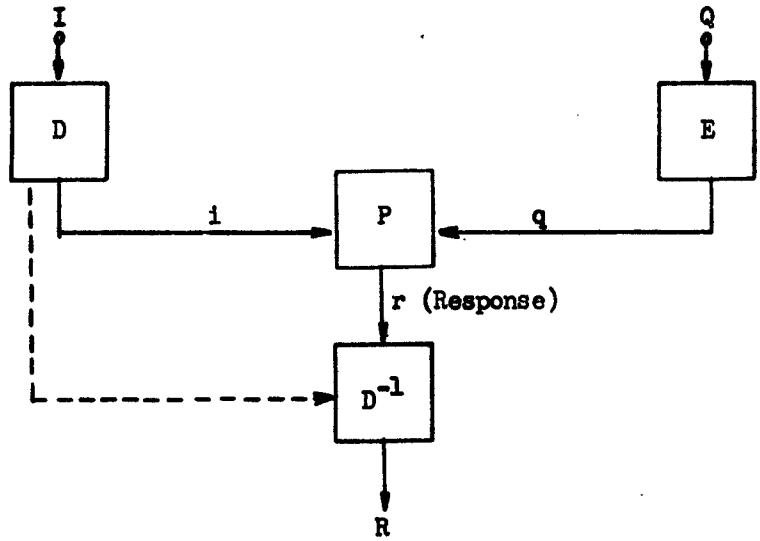


FIGURE 2. Model of General Retrieval System

retrieval systems as personnel files. But, more important, it is also valid for a general content retrieval system. This general retrieval model accepts input data, transforms the data into a convenient pre-established form, stores the data, and then selects responses from the data on the basis of questions or requests.

4.2.3 Analysis of the Model - It is now necessary to examine each functional section of the model represented by Figure 2 in order to analyze their differences and to determine the requirements for particular

types of systems.

4.2.3.1 The D Transform - For the literature search system, the D transform can be thought of as having two parts, D_d, and D_a. Symbolically, this form of the transform becomes:

$$D = D_d + D_a \tag{4-5}$$

The first part, D_d, can be viewed as a mapping of a docume.. .nto a set of symbols that represent the content of the document. A unique transformation D_d is generally not obtainable; this transformation varies with such intangibles as the analytic viewpoint and the number of descriptors assigned to a document. In the context of the literature search system the most that can be said of the transformation D_d is that, if the content descriptors i_d are processed by the processor p and the inverse transform D^{-1}, at least the original document will be obtained; however, many more documents, representing superfluous information, might also be obtained. Symbolically, this problem is written as:

$$D_d D_d^{-1} \neq 1 \tag{4-6}$$

On the other hand, the assignment of the address descriptor, D_a, is a unique transformation; it uniquely identifies a particular document. Symbolically:

$$D_a D_a^{-1} = 1 \tag{4-7}$$

Thus for the whole transformation D:

$$DD^{-1} = 1 \tag{4-8}$$

Descriptors are a restricted standard language. Documents and queries are transformed into this standard language, perhaps associated

with other standard terms, and then document identifiers are retrieved from the file. However, the number of descriptors and their richness could be gradually expanded by specifying them as actions, relations, results, means, or locations. At the same time, the constantly expanding descriptor language could be applied to sections of documents, then subsections, then paragraphs. A system can be postulated in which the descriptor language becomes as rich as the document or query itself--is in fact identical with the document or query--and in which the descripter language applies to units of information as small as sentences and phrases.

If the length of the descriptor list for each document is extended, a point will eventually be reached when each document is uniquely described; i.e., it would be possible, on the basis of a given set of descriptors, to select a particular docum t from the document store. At this point, $D_d D_d^{-1} = 1$; then D_d becomes redundant and can be eliminated. The point has not yet been reached where it is possible to stop storing documents and to rely upon their transformations (descriptors) and rules for their recreation from these transformations. When the descriptor list is extended, these descripters and an appropriate set of rules can recreate the desired document within certain limits. It may not be possible to obtain a word for word copy of the original, but the results will duplicate the meaning of the original.

Symbolically, this optimum system is represented by a slight modification to Equation 4-8:

$$DD^{-1} \doteq 1 \qquad (4\text{-}9)$$

This function means that, given a document I and the transformation algorithm D, a set of descriptors i can be obtained. Further transforming this set i by the inverse transformation D^{-1} does not yield I exactly, but it does yield a document I' that is close to I. Since i is a resultant of the transformation D(I)--see Equation 4-1, then:

$$I' = D^{-1}(i) = D^{-1}\left[D(I)\right] \qquad (4\text{-}10)$$

where $I' \triangleq I$. But the kind of transformation algorithm necessary for this system is the same as that necessary for content retrieval; any difference may be in the format or rules of the transformation.

It may be possible to transform a document for content retrieval into a format that is more compact and convenient than a list of descriptors. For content retrieval, each sentence would be transformed into a unique description that could readily be re-transformed into a close approximation of the original.

4.2.3.2 The P Transform - The major task of the P transform is to select sets of descriptors that have been stored in the system on the basis of their relationship to the request descriptors. Symbolically,

$$r = P(q,S_i) \qquad (4\text{-}11)$$

where r is the untransformed response. This transform can be viewed as having two parts; a storage function that stores and relates all the incoming descriptors and a selector function that matches the query descriptors, q, to the set stored descriptors, S_i.

For the literature search problem the selection process has criteria, among others, that should be noted:

(a) To maximize the amount of relevant information obtained.

(b) To minimize the number of irrelevant or erroneous answers.

For the content retrieval problem these criteria reduce to that of finding an acceptable answer to a query.

Thus, it is important for the transformation process P to be able to obtain or measure the degree of relevance of one set of information to another. These relationship indications are especially important for retrieval systems used for relatively uncategorized data in which many different descriptions of the same content might be possible--a condition that leads to difficulties in matching request and data descriptors. One way to indicate relationships among data is the kind of logical structure used to store information. The actual structure, however, may not be able to indicate the strength of these relationships; i.e., the degree of relevance or closeness among data. It may be necessary to provide a metric for the structure to determine the strengths of these relationships and to provide further indications of relevance, such as probabilistic measures, that may be incorporated into the storage structure. The selector function of P would use these relationships and their metrics as the basis of its selection algorithms. P may then be viewed as a combination memory, associational net, and selection mechanism.

In a literature search system the i and q would generally be descriptor lists, and P would store relationships between these descriptors. The output, r, of the transformation would be the identifying descriptors, usually addresses, of the relevant documents.

For a content retrieval system the i and q would have a more

complex format, but P would still be required to relate the various data elements to each other. The output, r, would be in the same format, as the i.

4.2.3.3 The E Transform - The E transform that transforms requests or queries into descriptor language is basically the same as the D transform. The major difference that might exist would be that of format; for the P transform might require a format for the transformed documents or input data. No address indication would be included in the transformed query, but the same kind of information would be indicated in the transformed query as in the transformed document. In other words, the same kind or similar language would be used for the transformed documents, i_d, and the transformed queries, q. Symbolically, these relations are:

$$\left.\begin{aligned} E &\sim D_d \qquad \text{(for literature search)} \\ E &\sim D \qquad \text{(for content retrieval)} \end{aligned}\right\} \quad (4\text{-}12)$$

The nature of these transforms is such that there is no loss of information in shifting from the i_d format to the q format and back; i.e.:

$$q = G(i_d) = G\left[G^{-1}(q)\right] \quad (4\text{-}13)$$

where G is the appropriate one-to-one transformation. In most cases G = 1, for q and i are expressed in the same format.

4.2.3.4 The D^{-1} Transform - For the literature search the D^{-1} transform is usually concerned with the addresses of documents. On the basis of these addresses the algorithm locates the documents in a file. The important part of this transform for the literature search is D_a^{-1},

which is an ever increasing file. If the system is a content retrieval system, then D^{-1} is not a file but a set of rules, comparable to D, for transforming the descriptor set, r, into the response, R.

4.2.4 Summary of the General Retrieval Model - The general information retrieval model can be summarized symbolically as follows. Given a set of documents or file items, S_I, in a retrieval system, T, a query Q produces a response R:

$$R = T\left[Q, S_I\right] \tag{4-14}$$

Some of the intermediate transformations that occur in this system can be written as:

$i = D(I)$ (the descriptor assigning process)

$q = E(Q)$ (the query transformation)

$r = P\left[q, S_i\right]$ (the selection process)

where D, E, and P are transformation algorithms; i, q, and r are input, query, and response descriptors; and S_i is the set of stored descriptors. Then:

$$R = D^{-1}\left\{P\left[q, S_i\right]\right\} \tag{4-15}$$

In terms of the original variables Q and I:

$$R = D^{-1}\left\{P\left[E(Q), S_{D(I)}\right]\right\} \tag{4-16}$$

since $S_i = S_{D(I)}$.

4.2.5 Specific Aspects of Retrieval Problem - Information retrieval systems, whether actual or theoretical, are composed of many elements. The general retrieval model highlighted three basic elements that any

usable information retrieval system must have:

(a) Descriptors or terms and their relationships.

(b) Files of data and/or terms or descriptors with an organization or structure.

(c) Procedures for searching files and locating data or terms.

Investigations into the problems associated with each of these areas are discussed briefly in the following paragraphs.

4.2.5.1 Descriptor Systems - Descriptors are introduced into information retrieval systems in order to reduce the language recognition and transformation requirements and to reduce the complexity of the data structures or content relationships. In short, descriptors represent an artificially restricted standard language to increase the convenience of handling requests, constructing and organizing the computer files, and searching for answers.

One of the major problems in constructing a descriptor system is the proper selection of the descriptors that are class names for synonyms so as to maximize retrieval of relevant information and minimize noise, the retrieval of irrelevant data. The descriptors must be words in common use, as unambiguous as possible, and sufficiently numerous to delineate relatively fine distinctions. Obviously, the more documents filed under a given descriptor, the larger the noise is likely to be.

To increase the number of relevant documents retrieved in response to a given request, descriptors for the request can be weighted. These weights can be assigned according to the relevance and the importance of the particular descriptor under consideration. The system can

then produce responses ordered according to weights assigned descriptors or responses greater than a fixed weight of relevance and importance. Another scheme for reducing irrelevance in responses is to assign descriptors to each section of documents added to the file. This method, of course, increases the degree of content retrieval.

Increasing the flexibility of descriptors by introducing role indicators or specifying terms as actions, relations, results, means, purposes, or locations is a further step toward content retrieval in the sense that it is the beginning of syntactical and semantic specification of request terms.

Some of the questions that must be answered before designing a descriptor system are:

(a) What descriptive terms are likely to be needed?

(b) How specific will the requests be?

(c) Will both specific and generic queries be made?

(d) Is the same information relevant to specific and generic queries?

(e) Is the correlation of the chosen descriptors sufficiently selective?

(f) If not, to what extent are interlocking, interfixing, and specifying of syntactic and semantic relations necessary and helpful?

4.2.5.2 Organization and Structure of Files - If information retrieval is viewed generally, it can be defined as locating and presenting a specific informative and accurate answer or piece of information in response to a specific question. Accomplishing this function requires

a classification scheme that groups larger units of related information; e.g., documents or sections of documents. Descriptors are assigned to units of information. The file consists of the system of descriptors and of information units ordered in some fashion to indicate the relations between descriptors and information. Generally, a descriptor is associated with many units of information and a unit of information may be described by several descriptors. In addition, the file structure must provide for relations among information units and among descriptors.

One of the best known systems that can be used to relate descriptors is the hierarchical classification or tree structure originally developed for biological classification. This type of structure forms a Boolean algebra under the relation of class inclusion. This type of model is appropriate only for a limited field of information in which a class is immediately subordinate to only one other class. This restriction requires a breakdown into small units of information, which means that the descriptor file would be composed of a large number of hierarchies of class inclusion. (The multilist system is a device for circumventing the limitations of ordinary list processing or hierarchies by allowing for relations among branches.)

For information fields of some diversity, the relations among descriptors usually form complicated networks to which the tree theory is not directly applicable. A general model of a complicated descriptor network is represented by means of a complemented modular lattice. This model is of sufficient generality to cover a wide variety of situations. Most elements are multiply connected rather than singly connected as in a

tree. The lattice model is referred to as a weak hierarchy--an element may have more than one predecessor. The tree is a strong hierarchy--an element has only one predecessor. The principal problem with the lattice model is that the number of nodes in the network quickly reaches into the millions if all relations between descriptors are represented. Consequently, the problem becomes one of effectively limiting the number of relations represented among descriptors.

The descriptor file associates descriptors with information units or items of data. These associations can be represented by a matrix of ones and zeros, where descriptors may be ordered as rows and information units as columns. A one indicates a relation; a zero, none. For a rich information store, this matrix will be large and most of its elements will be zeros. It is, therefore, an uneconomical representation. The matrix can be compressed by listing rows or columns (descriptors or data) and related items only for each entry. Of course, access to the file is much simpler for descriptor entry. Search time for these types of files can be reduced by using multiple entry of terms or by an ordered arrangement of both descriptors and data. Generic relations among terms can be shown by direct cross references, carried with each descriptor, or by a code of hierarchical class numbers showing the generic structure of the terms.

4.2.5.3 Search Procedures - In a retrieval system based upon descriptors there are two requirements for effective search. The first is the transformation of the request into the standard search terms. The second is the particular strategy or methodology for searching the descriptor

file effectively and fruitfully.

Transforming a request into standard descriptor terms is basically a form of translation from a rich language into a summary language or the matching of two sets of terms, one large, the other smaller. In order to accomplish this transformation, the meaning and relations between terms of the two sets or languages must be understood. Aid may be provided in the form of a dictionary or glossary of subject matter. The knowledge required to transform requests into descriptors is most simply provided to a computer by furnishing it with a thesaurus. Any more sophisticated means would involve a considerable capability for linguistic transformation on the part of the computer.

The formulation of a query and its transformation into a limited set of descriptors often does not provide sufficient information and direction to obtain exhaustive information concerning a subject that may exist in the data file. Effective search procedures are closely related to the way in which the descriptor file is structured and what sort of relations are indicated there. The most common method of searching is the conjunctive search, which retrieves only that information related to or encompassed by all the request descriptors in conjunction. There is a real need for investigating search procedures in terms of logical sums, differences, complements, and more complicated combinations of these functions as well as weighted logical functions in terms of set densities.

4.3 MEASURES OF RELEVANCE

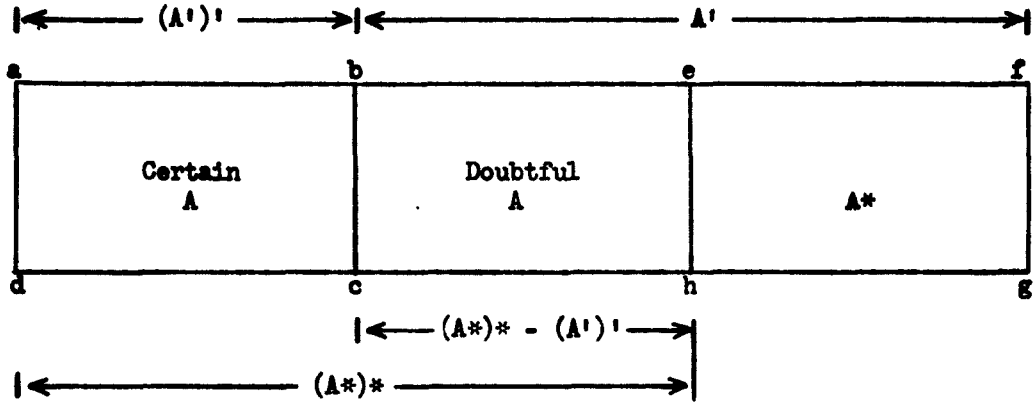
4.3.1 General - The formulation of a query and its transformation

into a limited set of descriptors often does not provide sufficient information and direction to obtain exhaustive information concerning a subject that may exist in the data file. Effective search procedures are closely related to the way in which the descriptor file is structured and to the sort of relations indicated by the structure. An effective information retrieval system must have the automated capability to associate other descriptors in the system, which are applicable or relevant to the topic in some degree, with those derived directly from the request. Several ways of determining the degree of dependence or relevance among descriptors have been suggested. Since this problem is a key aspect of information retrieval, some of the schemes for measuring the association or the relevance of terms are outlined and discussed briefly in the following paragraphs. These schemes are also reduced to a common system of notation to facilitate comparison.

4.3.2 Method 1 - This method is based upon the work of Fairthorne (1). Consider a set of items that has been completely classified or categorized under subject headings; that is, each item has been assigned to one or more categories. These items form a Boolean algebra in which the double complement law is valid. That is, the set of items that are not not-A's is identical with the set of all items that are A's, where A is a category. In a dynamic system, there will generally be items that have not been so classified, but knowledge of their existence would be helpful to the user. These items may not have been classified for several reasons: their proper classification is doubtful or unknown; they are not accessible; or, perhaps, there has been insufficient time to categorize them.

These items are now added to all the categories that might be relevant, including all the existing categories if relevance is completely unknown. With this classification scheme, all but not only or only but not all items can be retrieved--the first by including items in the doubtful category, the second by ignoring items in the doubtful category.

This concept can be expressed more formally. If the correct classification of some items is doubtful, a system has two types of complements of a given set of terms. These complements comprise the inclusive and exclusive complements of sets as shown in Figure 3. The set A is represented



LEGEND: A = Set under consideration
A* = Exclusive complement of A
(A*)* = Exclusive complement of A* (all but not only)
A' = Inclusive complement of A
(A')' = Inclusive complement of A' (only but not all)
(A*)* - (A')' = Doubtful A

FIGURE 3. Inclusive and Exclusive Complements of Sets

by the rectangle abcd plus an a priori unknown number of documents in

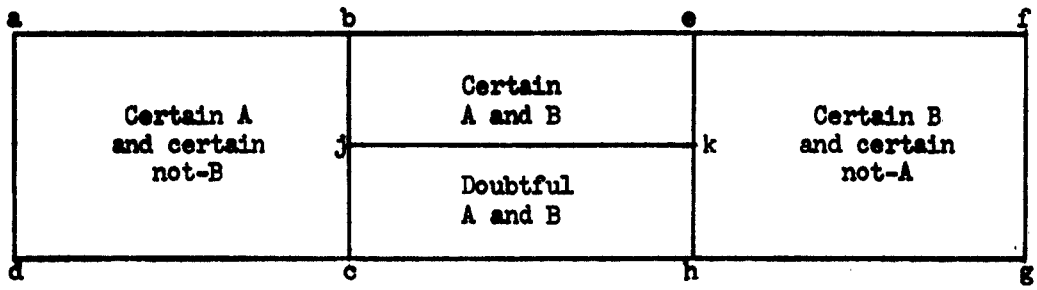
the rectangle behc. The exclusive complement A* of a set A is defined as the largest set of items that certainly does not contain any members of A; A* is represented by the rectangle efgh. Then, (A*)* is the smallest set of items that certainly contains all the members of A; namely, the rectangle aehd. The inclusive complement A' is defined as the smallest set of items that certainly contains all the items that are not members of A; clearly, this set is the rectangle bfgc. (A')' is the largest set that certainly contains only elements of A. Thus (A')' is the rectangle abcd. Documents of ambiguous or doubtful classification will be elements of (A*)*. When their proper classification has been resolved, they become elements of (A')'.

Define the distance, d, between two sets as the number of elements in their symmetric difference. That is:

$$d(A,B) = A-B \cup B-A \qquad (4\text{-}17)$$

This definition has the properties that a good definition of distance should have. In particular, it satisfies the axioms for distance in a metric space.

This concept can be applied to the classification scheme just discussed. The interpretation of distance in this case is the remoteness or irrelevance of two topics. There are two distances corresponding to the two complements, as illustrated by Figure 4. The inclusive distance is the least set of items that certainly includes all items that belong to one but not both of the sets. This set is represented by the sum of the rectangles abcd, efgh, and jkhc in Figure 4. The exclusive distance is the largest set of items that certainly belongs to one of the sets



LEGEND: Inclusive distance between A and B: abcd + efgh + jkhc
Exclusive distance between A and B: abcd + efgh
Measure of uncertainty of relevance of A and B: jkhc

FIGURE 4. Inclusive and Exclusive Distance Between Sets

but not both; that is, the rectangles abcd plus efgh. Obviously, the inclusive distance is always greater than or equal to the exclusive distance. The difference between the two distances--namely, the rectangle jkhc--measures the current uncertainty about the relevance of the two topics in a particular system. Documents of uncertain classification are in the set $(A^*)^* - (A')'$. This point is evident in Figure 3.

4.3.3 Method 2 - A second measure of distance between topics is adapted from Klingbiel (5) and is a modification of the first. This measure is a normalized version of Equation 4-17. Method 1 produces inordinately large distances for large sets. The purpose of the second method is to obtain a measure that is more independent of the number of elements in the set. The modified definition is:

$$d(A,B) = \frac{A \cup B - A \cap B}{A \cap B} = \frac{A \cup B}{A \cap B} - 1 \qquad (4\text{-}18)$$

4.3.4 Method 3 - The third method is adapted from Mooers (8). Information concerning a given topic can be thought of as a conjunction of applicable descriptors. The closeness of two topics can be measured by a comparison of weighted descriptors that the two topics have in common. The descriptors of the system can be identified with an ordered sequence of binary bits. Each descriptor is represented by a position in the binary number. If the descriptor is applicable to a certain topic A, then a 1 appears in that position, otherwise a 0. Each position is also assigned a postive weight, w_j (for the $j^{\underline{th}}$ bit), indicating the importance or degree of relevance of that descriptor to the topic. The distance d between two topics A and B can then be defined as:

$$d(A,B) = \frac{[(\Sigma w_j a_j)(\Sigma w_j b_j)]^{\frac{1}{2}}}{\Sigma w_i a_j b_j} - 1 \qquad (4\text{-}19)$$

where a_j and b_j are the $j^{\underline{th}}$ bits of the respective ordered descriptor numbers. This definition requires that there be at least one descriptor common to the two topics. An anomaly of this definition is that it does not satisfy the axioms of distance in a metric space. In particular, it is not necessarily the case that $d(A,C) \leq d(A,B) + d(B,C)$.

4.3.5 Method 4 - This method, which has been discussed by Watanabe (10), considers a probabilistic model for the association of terms. It associates either descriptors or items on the basis of the correlation among them. The relationship between items and descriptors is presented in the form of a matrix. In this matrix each element represents the assignment or non-assignment of a descriptor to an item. The item-descriptor matrix, T, is then defined as an m by n matrix whose element

$T(x_i, y_j)$ of the $i^{\underline{th}}$ row and $j^{\underline{th}}$ column is 1 or 0, according to whether item x_i does or does not have descriptor y_j.

Consider now a large collection of items, $X = (x_i)$, $i = 1,2,\ldots,m$, with a set of descriptors, $Y = (y_j)$, $j = 1,2,\ldots,n$. The probability that an arbitrary item has the description $b_1,b_2\ldots,b_n$, which is an ordered sequence of bits representing the applicability of the n descriptors, is the ratio of the number of rows with the proper bit pattern to the total number of rows in the matrix. This probability is expressed by:

$$p(b_1,b_2,\ldots,b_n) = p(Y) = \sum_{i=1}^{m} \frac{1}{m} \prod_{j=1}^{n} \delta[b_j, T(x_i, y_j)] \tag{4-20}$$

where δ is the Kronecker delta, so that $\delta(a, b) = 0$ if $a \neq b$, and $\delta(a, b) = 1$ if $a = b$.

For the collection of items the uncertainty about the description of an arbitrarily selected object can be measured by an entropy function, $S(Y)$:

$$S(Y) = -\Sigma p(Y) \log p(Y) \tag{4-21}$$

where the summation is extended over two values, 0 and 1, for all the b's corresponding to Y. Similar entropy functions can be defined over subsets of descriptors, Y_μ, such that:

$$S(Y_\mu) = -\Sigma p(Y_\mu) \log p(Y_\mu) \tag{4-22}$$

with the summation extending over the two values of all the b's corresponding to Y_μ.

An information theoretical measure of correlation can be defined for a set of descriptors Y_μ with respect to its disjoint subsets $Y_{\mu\sigma}$ by:

$$C(Y_\mu;\ Y_{\mu 1},\ Y_{\mu 2},\ldots,Y_{\mu p}) = \sum_{\sigma=1}^{\varphi} S(Y_{\mu\sigma}) - S(Y_\mu) \qquad (4\text{-}23)$$

where the $Y_{\mu\sigma}$ are disjoint and complete subsets of Y_μ so that any element, y_j, of Y_μ belongs to one and only one of these subsets. The correlation, C, may be considered as a generalization of the information function.

The total correlation in Y_μ, $C_T(Y_\mu)$, can be considered as the redundancy existing in Y_μ among its elements, $y_j \in Y_\mu$. The total correlation then is:

$$C_T(Y_\mu) = \sum_{y_j \in Y_\mu} S(y_j) - S(Y_\mu) \qquad (4\text{-}24)$$

Of course, the total correlation in Y is simply:

$$C_T(Y) = \sum_j S(y_j) - S(Y) \qquad (4\text{-}25)$$

$C_T(Y)$ is the largest of all possible $C(Y;\ Y_1,\ Y_2,\ldots,Y_\theta)$.

Correlation between two descriptors, y_k and y_r, can be broken into two parts, similarity and dissimilarity:

$$C(y_k,\ y_r) = C^+(y_k,\ y_r) + C^-(y_k,\ y_r) \qquad (4\text{-}26)$$

Similarity, $C^+(y_k,\ y_r)$, is a measure of the number of times y_k and y_r are jointly assigned or not-assigned to the same item. Dissimilarity, $C^-(y_k,\ y_r)$, is a measure of the number of times y_k and y_r are oppositely assigned to the same item. Similarity can then be expressed by:

$$C^+(y_k,\ y_r) = \sum_{b_k, b_r} \delta[b_k,\ b_r]p(b_k,\ b_r) \log \frac{p(b_k,\ b_r)}{p(b_k)p(b_r)} \qquad (4\text{-}27)$$

And dissimilarity can be expressed by:

$$C^{-}(y_k, y_r) = \sum_{b_k, b_r} \delta[b_k, 1 - b_r] p(b_k, b_r) \log \frac{p(b_k, b_r)}{p(b_k)p(b_r)} \quad (4\text{-}28)$$

In this description only the correlation between descriptors has been indicated. If, however, items are considered instead of descriptors, the correlation, similarity, and dissimilarity of objects may be measured by the same formulae.

4.3.6 Utility of Measures - The utility of these measures of association, distance, and similarity lies in the fact that they provide an automatic means of relating request descriptors to other descriptors and relating documents to other documents or information. For example, a request descriptor could be given and the system would be asked to retrieve all information under descriptors with a similarity to the given descriptor (in the sense of Method 4) greater than some prescribed number. This process can appropriately be called concept retrieval. Note that concept retrieval can be applied to either content retrieval or document (partial content) retrieval. This notion of concept must possess a kind of continuity, namely that a small change in the set of objects under consideration must produce only a small change in the concept. Otherwise, the definition is clearly not in accord with intuition. The other definitions of distance can be used in a similar fashion to assist in obtaining relevant descriptors and/or to retrieve information ordered according to relevance.

The measures outlined here will not be evaluated further in this report except to state that two types of evaluation are possible. The first is the theoretical adequacy of a definition and its implications.

The second is the ultimate evaluation test, namely the utility of the definitions in terms of actual use in retrieving information in an operational information retrieval system. That is, does the concept in practice effectively assist in the retrieval of information judged to be relevant to the request by the requestor?

4.4 REFERENCES

(1) Fairthorne, R. A.; "Delegation of Classification," American Documentation; Volume 9, March 1953.

(2) Gray, H. J., et al; Information Retrieval and the Design of More Intelligent Machines; Final Report No. AD59URI to the U. S. Signal Corps, July 1959.

(3) Gray, H. J., et al; The Multi-List System; Report to the Office of Naval Research, Information Systems Branch, under Contract NOnr551(40), November 1961.

(4) Green, B. F., Wolf, A. K., Chomsky, Carol, and Laughery, K.; "Baseball: An Automatic Question-Answer," Proceedings WJCC; IRE, Los Angeles, May 1961.

(5) Klingbiel, P. H.; Language Oriented Retrieval Systems, (AD 271-600); February 1962.

(6) Luhn, H. P.; Auto-Encoding of Documents for Information Retrieval Systems; IBM Research Center, Yorktown Heights, New York, 1958: 7 pp.

(7) Maron, M. E.; Automatic Indexing: An Experimental Inquiry, (AD 245-175); RAND Corporation, Santa Monica, California, 10 August 1960: 37 pp.

(8) Mooers, C. N.; The Use of Symbols in Information Retrieval, RADC-TN-59-133; (AD 213-782); April 1959.

(9) Perry, J. W., Kent, A., and Berry, M. M.; Machine Literature Searching; New York, 1956.

(10) Watanabe, S.; A Probabilistic View of the Formation of Concept and of Association; presented at the annual meeting of the AAAS, 26-30 December 1961.

(11) Personal communications and informal briefing.

5. CONCLUSIONS

Four aspects of the research orientation have been described: system-procedure, real-hypothetical, hardware-software, reduction-manipulation. A theoretical--procedural, hypothetical, software, manipulative--approach is being taken. A preliminary generalized model has been formulated, and some of its implications have been considered. One procedural area, the measurement of relevance, has been formally elaborated. Further work on the functional characteristics of a general theory of information retrieval, the development of the model, and the formal consideration of additional procedures and techniques is required.

6. PLANS FOR THE NEXT QUARTER

Activities during the next quarter will proceed with the over-all goal of developing a theory of information retrieval for use as a tool in the design of information retrieval systems. Work will include at least the following three aspects of the development of such a theory.

(a) A statement of the necessary or desirable features of a theory of information retrieval together with a breakdown of the essential functional elements of information retrieval and their interrelationships.

(b) Continue development of an information retrieval model based upon Item (a) and the models described in this report. This work will use and relate the results of Item (c).

(c) Continue work on functional elements of the model and techniques that are applicable to the effective performance of these essential functions (e.g., measures of relevance as applied to descriptor assignment).

These three aspects of the work are actually levels of detail. The first provides a general statement of the objectives of the research, defines essential areas of effort, and provides guidelines and definitions for use in the development of the theory. The second level of effort develops and defines the essential features of the theory to the point where a representative model is meaningful. It will isolate independent functions and establish relations between functions that are not independent. The third level develops detailed techniques, procedures, and methodology useful for the design of an effective information retrieval system.

7. IDENTIFICATION OF PERSONNEL

7.1 PERSONNEL ASSIGNMENTS

The following personnel were assigned to the project during the period covered by this report:

Name	Title	Man-Hours
Jacques Harlow	Manager	50
Quentin A. Darmstadt	Research Specialist	300
George Greenberg	Senior Specialist	350
Alfred Trachtenberg	Senior Program Analyst	550

7.2 BACKGROUND OF PERSONNEL

7.2.1 Jacques Harlow - AB, Philosophy, Dartmouth College, 1950; PhD candidate, statistics and economics, New York University, 1963. Manager of basic and applied research activities oriented to new uses of electronic digital computers. Activities include problem-oriented languages, man-machine communication, models of artificial intelligence, adaptive control processes, and linguistic analysis.

7.2.2 Alfred Trachtenberg - BS, Electrical Engineering, Columbia University, 1956; MS, Electrical Engineering, Columbia, 1958; Degree in Electrical Engineering, Columbia, 1962. Activities center on the development of a model of learning that is applicable to non-biological systems. Previous experience includes the analysis, evaluation, and design of complex radar, control, and defense systems.

7.2.3 Quentin A. Darmstadt - AB, Mathematics, Oberlin College, 1950; advanced studies in mathematics and mathematical logic, Harvard University,

and New York University. Activities center upon developing logical and mathematical proofs leading to the formulation of algorithms for solving problems on electronic digital computers. Experience includes operational analysis and evaluation of systems.

7.2.4 George Greenberg - BA, Psychology, Brooklyn College, 1955; PhD, Psychology, Duke University, 1960. Activities include psychological research in learning, psycho-linguistics, and perception. Previous experience includes the organization of research into the automation of command languages.

DISTRIBUTION LIST

Recipient	Copies
OASD (R&E) Rm 3E1065 Attention: Technical Library The Pentagon Washington 25, D. C.	1
Chief of Research and Development OCS, Department of the Army Washington 25, D. C.	1
Commanding General U. S. Army Materiel Command Attention: R&D Directorate, Res Div, Elect Br. Washington 25, D. C.	1
Commanding General U. S. Army Electronics Command Attention: AMSEL-AD Fort Monmouth, New Jersey	3
Commander, Armed Services Technical Information Agency Attention: TIPOR Arlington Hall Station Arlington 12, Virginia (Reports)	10
Commanding General USA Combat Developments Command Attention: CDCMR-E Fort Belvoir, Virginia	1
Commanding Officer USA Communication and Electronics Combat Development Agency Fort Huachuca, Arizona	1
Commanding General U. S. Army Electronics Research and Development Activity Attention: Technical Library Fort Huachuca, Arizona	1

Recipient	Copies
Chief, U. S. Army Security Agency Arlington Hall Station Arlington 12, Virginia	2
Deputy President U. S. Army Security Agency Board Arlington Hall Station Arlington 12, Virginia	1
Director, U. S. Naval Research Laboratory Attention: Code 2027 Washington 25, D. C.	1
Commanding Officer and Director U. S. Navy Electronics Laboratory San Diego 52, California	1
Aeronautical Systems Division Attention: ASAPRL Wright-Patterson Air Force Base, Ohio	1
Air Force Cambridge Research Laboratories Attention: CRZC L. G. Hanscom Field Bedford, Massachusetts	1
Air Force Cambridge Research Laboratories Attention: CRXL-R L. G. Hanscom Field Bedford, Massachusetts	1
Headquarters, Electronic Systems Division Attention: ESAT L. G. Hanscom Field Bedford, Massachusetts	1
Rome Air Development Center Attention: RAALD Griffiss Air Force Base, New York	1

Recipient	Copies
AFSC Scientific/Technical Liaison Office U. S. Naval Air Development Center Johnsville, Pennsylvania	1
Commanding Officer U. S. Army Electronics Materiel Support Agency Attention: SELMS-ADJ Fort Monmouth, New Jersey	1
Director, Fort Monmouth, Office USA Communication and Electronics Combat Development Agency Fort Monmouth, New Jersey	1
Corps of Engineers Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Marine Corps Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
AFSC Scientific/Technical Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Logistics Division Fort Monmouth, New Jersey Attention: Anthony V. Campi	9
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Director of Research/Engineering Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Technical Documents Center Fort Monmouth, New Jersey	2

Recipient	Copies
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: SELRA/NPE Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Technical Information Division Fort Monmouth, New Jersey	3
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Exploratory Research Dr. Reilly Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Engineering Sciences Department Mr. Hennessy Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Exploratory Research Jack Benson Fort Monmouth, New Jersey	3